## Abstract

The following paper reproduces a multiple linear regression model used in *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. This book is avaialbe [ENTER URL]/The regression model pertains to predicting **Sales** (in the thousands) from the advertising budgets of three different predictors: **TV**, **Radio**, and **Newspaper** . It is featured in chapter 3.2 of the book.

## Introduction

Given a data set containing information pertaining to both sales and advertising budget, how could one develop a marketing plan that would result in higher product sales?

To achieve this goal, the following paper will display how improving sales via multiple linear regression can be implemented. We will attempt to model the sales of a particular product as a function of advertising budget for three types of media: TV, radio, and newspaper. Because we are modelling this relationship via linear regression, we are inherently assuming that a linear relationship exists between our dependent variables and independent variable. We will run a total of four regressions: one for each of the sale + predictor combinations, and one where we regress sales on all three predictors.

## Data

The data set utilized in this paper, **Advertising.csv**, contains 200 observations of 5 variables. Four of the variables are of general interest: **Sales**, **TV**, **Radio**, and **Newspaper**. The fifth variable, **X**, is an index.

This dataset is free to access online at the following url [URL]. Alternatively, you can visit the author's GitHub repository My Github Account to access it as well.

For this paper, we will use the following variables: **Sales**, **Radio**, _Newspaper___, and **TV**. **Sales** refers to the number (in thousands) of a particular product's sales in 200 different markets. The other three variables correspond to advertising budget for each medium (in the thousands of $)

In our analysis, we will treat **Sales** as the dependent variable and **TV**, **Radio**, and **Newspaper** as the independent variables.

# Methodology

### Multiple Linear Regression

As stated, our approach for determining the association between **Sales** and the other three variables,**Radio**, **Newspaper**, and **TV**, will be linear regression. In particular, we will use simple linear regression: predicting **Y** (a continuous, quantitative response) on the basis of single predictor variables, **X** (where **X** is a set of three variables). We'll assume the following linear relationship, regressing Y onto X:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Our B parameters are unkown constants that represent the how the dependent variable changes given a one-unit chane in the corresponding variable. These parameters are known as the model *coefficients*. By fitting a linear model onto our training data, we obtain coefficient estimates that we use in the aforementioned linear model. In order to get our line, we must estimate the regression coefficients.

### Estimating the Regression Coefficients

The parameters are estimated using the same least squares approach that we used for simple linear regression; i.e. choose B0, B1,. . . ,Bp to minimize the sum of squared residuals (RSS):

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

The values B^0, B^1,. . . , B^p that minimize the above RSS equation are the multiple least squares regression coefficient estimates. These are obtained via linear algebra operations.

Unlike the simple linear regression estimates given in (3.4), the multiple regression coefficient estimates have somewhat complicated forms that are most easily represented using matrix algebra. For this reason, we do not provide them here. Any statistical software package can be used to compute these coefficient estimates, and later in this chapter we

## Important Questions in Multiple Regression

When we perform multiple linear regression, we usually are interested in answering the following questions:

1. Is at least one of the predictors X1, X2,. . . ,Xp useful in predicting the response?
2. Do all the predictors help to explain Y , or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Below we'll discuss them piecewise.

### 1. Is at least one of the predictors X1, X2,. . . ,Xp useful in predicting

the response?

Recall, for the simple regression case, we could test whether X1 was associated with Y via a simple t-test with the null hypothesis dictating that B1 = 0. This won't work for multiple regression, because we're dealing with multiple predictors.

In the multiple regression case, say with p predictors, we need to test whether all of the regression coefficients are zero, (i.e. whether B1 = B2 = · · · = Bp = 0).

We'll use a hpyothesis test to achieve this.

$$H_0 : \beta_1 = \beta_2 = ...\beta_p = 0$$
$$H_1 : \exists \beta_j \neq 0$$

H0 : B1 = B2 = · · · = Bp = 0 versus the alternative Ha : at least one Bj is non-zero. This hypothesis test is performed by computing the F-statistic:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

### 2. Deciding on Important Variables

If our F-test reveals that at least one of the predictors is related to the response, we can investigate which variables are related.

There are numerous methods to achieve this. If we have a small number of features, we can investigate the p-values one-by-one. However, this is not feasible for large datasets with high-dimensionality. The task of determining which predictors are associated with the response is referred to as variable selection, and our discussion will be limited to only a few classical approaches to vairable selection.

One method of variable selection is to try each model corresponding to all the combination of coefficients that we have. When selecting a model this way, multiple assessment criteria exist, such as *Mallows Cp, Akaike information criterion (AIC), Bayesian informationcriterion (BIC)*, and *adjusted R-squared.*

Determining all possible model forms is not trivial. There are a total of $2^p$ models that contain subsets of p variables; this grows exponentially and quickly makes the task of trying out every possible subset of the predictors infeasible. Therefore, unless p is very small, we cannot consider all $2^p$ models, and instead we'll utilize three automated approaches to choose a smaller set of models.

#### Forward selection.

In forward selection, we begin with a model that only has the intercept terms. From this, we fit all p possible simple linear regression models and add the variable that results in the lowest $RSS$ score. Then, with that model, we consider all other $p-1$ variables and add the one that results lowest $RSS$ for the two variable model. We repeat this until some stopping criteria is met.

#### Backward selection

Backward selection is similar to forward selection, except that instead of starting with the null model, we start with all variables. From there, we iteratively remove the variable with the largest p-value (as that variable is the least statistically significant). We then fit the new model and repeat until some stopping criteria is met (e.g. all p-values below some threshold).

#### Mixed selection

This is a combination of forward and backward se- mixed lection. We start with no variables in the model, and as with forward selection selection, we add the variable that provides the best fit. As we add variables one-by-one, the p-values may become larger. Thus, if any p-value rises above a certain threshold during the variable addition, we remove that variable from the model. We continue to perform these forward and backward steps in this manner until all variables in the model have a sufficiently low p-value, and all variables outside the model have large p-values (when added to the model).

All of these methods have best-case uses, but it should be noted that backward selection cannot be used if $p > n$ , while forward selection can always be used. Also forward selection is a greedy approach.

### 3. Model Fit

Two common numerical measures of model fit we will discuss are the $RSE$ and $R^2$, the fraction of variance explained.

R-squared

$R2$ is the square of the correlation of the response and the variable. In multiple linear regression, this equals $Cor(Y, Y\hat{})^2$, the square of the correlation between the response and the fitted linear model. A notable property of the fitted linear model is that it maximizes this correlation among all possible linear models.

A $R^2$ close to 1 indicates a good fit, while a value near 0 indicates a poor fit (though it's problem dependent).

One caveat regarding $R^2$ is that it will always increase when more variables are added to the model, even if those variables are only weakly associated with the response. This occurs because adding another variable to the least squares equations must allow us to fit the training data (though not necessarily the testing data) more accurately. Thus, the $R^2$, which is also computed on the training data, must increase, thereby overfitting the training set.

### RSE

RSE is the residual standard error. It is an estimate of the standard deviation of the error in our model. Intuitively, it is the average amount that our response will deviate from the true regression line, defined as follows:

$$\sqrt{1/(n-2) * \sum_{i=1}^{n}(y_i - \hat{y_i})^2}$$

We consider this a measure of the *lack of fit* of our model and want it as small as possible.

Finally, a simple method to assess model fit is to plot the data.

## 4. Predictions

Once we have fit the multiple regression model, applying it for prediction is farily starightforward. That said, we must be cautious of three things when making preditions with our model:

1 1. The coefficient estimates (our $\hat{(}\beta)$ terms are estimate for the *true* $\beta$ values. As such, our least squares plan is an approximation to the true population regression plane. We can compute a confidence interval in order to determine how close $Y$ will be to $f(X)$.

2. There is an additional source of potentially reducible error: model bias. Even if our approximate model were correct, this discrepancy will always exist. Regularization parameters exist to help take care of this.

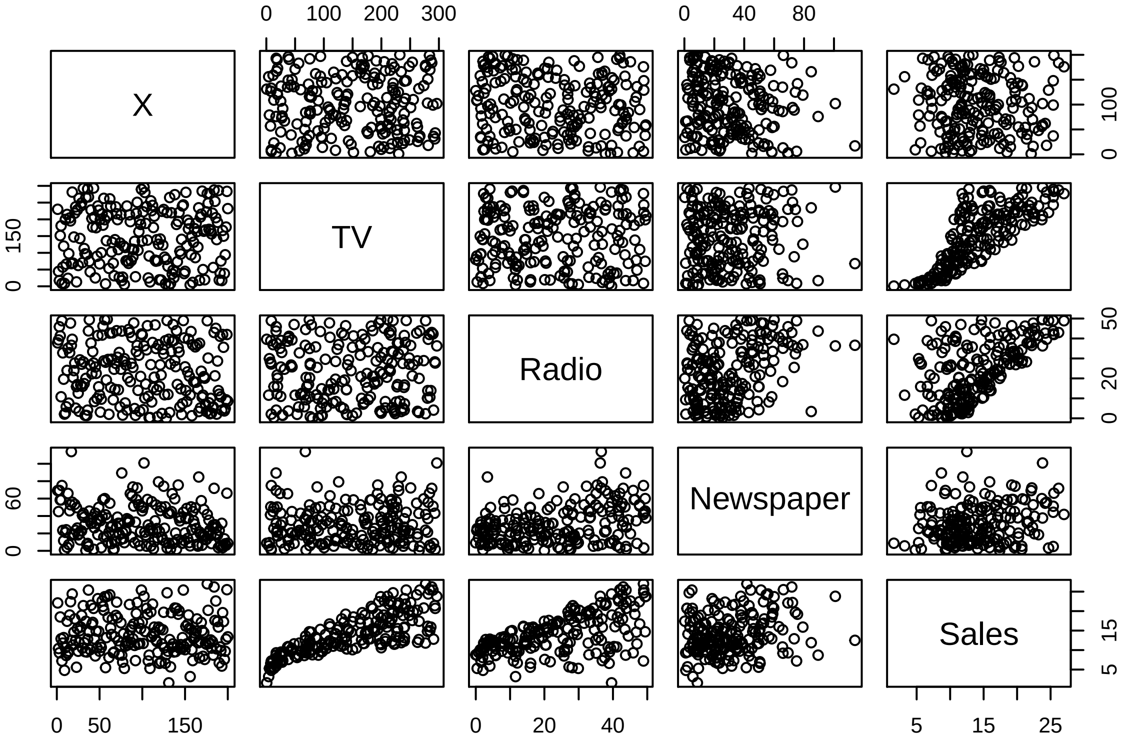
3. There is also random error in the model (aka irreducible error). We can assess the discrepancy betwween our predicted response and the true response using prediction intervals. Note that prediction intervals are always wider than confidence intervals because the incorporate the the estimation error *and* the uncertainty of our irreducible error.

## Results

This section will apply all we learned in the above methodology section to our the *Advertising* data set.

To begin, let's view some of our data. We'll begin by viewing a scatterplot for each variable.

```r
adv <- read.csv("/Users/jared/Desktop/stat159/stat_159/stat159-fall2016-hw03/data/Advertising.csv")
#source("code/scripts/regression-script.R")
library(xtable)
lm.fit <- lm(Sales ~ TV + Radio + Newspaper, data=adv)
lmsum <- summary(lm.fit)
pairs(adv)
```



Our regression problem assumes the following form:

In our case, the model appears as follows:

$$sales = \beta_0 + \beta_1 * TV + \beta_2 * Radio + \beta_3 * Newspaper + error$$

How correlated are our features?

% latex table generated in R 3.3.1 by xtable 1.8-2 package % Fri Oct 14 22:37:16 2016

| | X | TV | Radio | Newspaper | Sales |
|---|---|---|---|---|---|
| X | 1.00 | 0.02 | -0.11 | -0.15 | -0.05 |
| TV | 0.02 | 1.00 | 0.05 | 0.06 | 0.78 |
| Radio | -0.11 | 0.05 | 1.00 | 0.35 | 0.58 |
| Newspaper | -0.15 | 0.06 | 0.35 | 1.00 | 0.23 |
| Sales | -0.05 | 0.78 | 0.58 | 0.23 | 1.00 |

What happens if we fit individual regression models (one for each independent variable)?

% latex table generated in R 3.3.1 by xtable 1.8-2 package % Fri Oct 14 22:37:16 2016

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 7.0326 | 0.4578 | 15.36 | 0.0000 |
| TV | 0.0475 | 0.0027 | 17.67 | 0.0000 |

Table 1: Sales on TV

% latex table generated in R 3.3.1 by xtable 1.8-2 package % Fri Oct 14 22:37:16 2016

% latex table generated in R 3.3.1 by xtable 1.8-2 package % Fri Oct 14 22:37:16 2016

Unfortunately, using multiple linear models makes it difficult to estimate response values. Such a method also leaves room for surrogating and neglects confounding factors, such as influence from other variables.

|  | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 12.3514 | 0.6214 | 19.88 | 0.0000 |
| Newspaper | 0.0547 | 0.0166 | 3.30 | 0.0011 |

Table 2: Sales on Newspaper

|  | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 9.3116 | 0.5629 | 16.54 | 0.0000 |
| Radio | 0.2025 | 0.0204 | 9.92 | 0.0000 |

Table 3: Sales on Radio

For this reason, we will use a multiple linear regression model:

% latex table generated in R 3.3.1 by xtable 1.8-2 package % Fri Oct 14 22:37:16 2016

|  | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.0052 | 0.3942 | 7.62 | 0.0000 |
| X | -0.0006 | 0.0021 | -0.28 | 0.7827 |
| TV | 0.0458 | 0.0014 | 32.73 | 0.0000 |
| Radio | 0.1884 | 0.0086 | 21.78 | 0.0000 |
| Newspaper | -0.0012 | 0.0059 | -0.21 | 0.8342 |

Table 4: Multiple Linear Regression for Advertising data set

Thus we can see the true effect for each variable; namely that TV and Radio are significant predictors but Newspaper is not. Thus, we should probably decrease our budget in newspaper ads and move it into TV and Radio.

Now that we've created our model, how does it fit (or fail to fit) our data? To assess the performance of our multiple linear regression model, we'll employ the methods we discussed previously.

```
## % latex table generated in R 3.3.1 by xtable 1.8-2 package
## % Fri Oct 14 22:37:16 2016
## \begin{table}[ht]
## \centering
## \begin{tabular}{lr}
##   \hline
## Quantity & Value \\
##   \hline
## Residual Standard Error & 1.68 \\
##   R-Squared & 0.90 \\
##   F-Statistic & 570.27 \\
##    \hline
## \end{tabular}
## \caption{Multiple Linear Regression Coefficients}
## \end{table}
```

Thus, our value for R-squared (.9) suggests that our model fit the data pretty well. Furthermore, our huge value for our F-statistic (570.27) reveals that this model is useful because a relationship definitely exists between our dependent and independent variables. Finally, our Residual Standard Error value (1.68) is not too large.

## Conclusion

In conclusion, multiple linear regression is a great choice of model when you have multiple predictors, a continuous response type, and assume a linear relationship between the response and predictors. In order to assess the performance of such a model, you can employ numerous techniques. We used r-squared, residual standard error, and the f-statistic. Our results for the advertising data set yielded that there is a significant relationship between Sales and two of our predictors: TV and Radio. Newspaper was insignificant.